How Well Can We Correct Systematic Errors in Historical XBT Data?

LIJING CHENG,a HAO LUO,a,b TIMOTHY BOYER,c REBECCA COWLEY,d JOHN ABRAHAM,e VIKTOR GOURETSKI,f FRANCO RESEGHETTI,g AND JIANG ZHUa,b

a International Center for Climate and Environment Sciences, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China
b University of Chinese Academy of Sciences, Beijing, China
c National Centers for Environmental Information, Silver Spring, Maryland
d Oceans and Atmosphere, CSIRO, Hobart, Tasmania, Australia
e University of St. Thomas, St. Paul, Minnesota
f Center for Earth System Research and Sustainability, University of Hamburg, Hamburg, Germany
g SSPT-PROTER-BES, Italian National Agency for New Technologies, Energy and Sustainable Economic Development, Lerici, Italy

(Manuscript received 11 July 2017, in final form 27 December 2017)

ABSTRACT

Biases have been identified in historical expendable bathythermograph (XBT) datasets, which are one of the major sources of uncertainty in the ocean subsurface database. More than 10 correction schemes were proposed; however, their performance has not been collectively evaluated and compared. This study quantifies how well 10 different available schemes can correct the historical XBT data by comparing the corrected XBT data with collocated reference data in both the World Ocean Database (WOD) 2013 and the EN4 dataset. Four different metrics are proposed to quantify their performances. The results indicate CH14 is the best among the currently available methods, and L09/G12/GR10 can be used with some caveats. To test the robustness of the schemes, we further train the CH14 and L09 by using 50% of the XBT–reference data and the schemes are tested by using the remaining data. The results indicate that the two schemes are robust. Moreover, the EN4 and WOD comparison datasets show a systematic difference of XBT error (−0.01°C on a global scale and 0–700 m on average). Influences of quality control and data processing have been investigated. Additionally, the side-by-side XBT–CTD comparison experiment is used to examine the correction schemes and provides independent high-quality data for the assessment. The schemes that best correct the global datasets do not always perform as well at correcting the side-by-side dataset, and further examination of the discrepancy in performance is still required. Finally, CH14 and L09 result in very similar ocean heat content (OHC) change estimates in the upper 700 m since 1966, suggesting the potential of reducing XBT-induced error in OHC estimates.

1. Introduction

Biases linked to expendable bathythermograph (XBT) observations are one of the major sources of error in ocean subsurface temperature datasets (Lyman et al. 2010; Abraham et al. 2013; Rhein et al. 2013; Goes et al. 2015a,b; Cheng et al. 2016; Boyer et al. 2016), introducing uncertainty of changes in ocean heat content (Abraham et al. 2013; Rhein et al. 2013). Therefore, a dataset with instrumental biases minimized would contribute to a more accurate detection of global ocean change, including uncertainty estimation and attribution to natural and anthropogenic drivers.

Decades of efforts have been made by the XBT community to identify and understand the cause of errors in XBT measurements and to quantify its magnitude. In brief, the genesis of XBT errors stem from their initial purpose: to make lower-quality measurements primarily for submarine operations. Later, temperature measurements from XBT devices were adopted for oceanographic purposes and the accuracy/depth requirements for this purpose are more stringent. The errors stem from the lack of a depth sensor on board the XBT device; its depth is estimated from approximating...
functions whose independent variable is time. The approximating function is generally used without regard to manufacturing changes or tolerances, changes to drop height, differences in water temperature, and other operational variables. A summary of these issues is provided in Abraham et al. (2013). Since the 1970s, many analyses on side-by-side XBT and conductivity–temperature–depth (CTD) comparison experiments have been constructed to quantify the XBT errors (Flierl and Robinson 1977; Anderson 1980; Green 1984; Hallock and Teague 1992; Bailey et al. 1994; Hanawa et al. 1995; Reseghetti et al. 2007; Cowley et al. 2013). These experiments were performed by near-simultaneous deployment of XBT and CTD [salinity–temperature–depth (STD) in earlier tests] devices from near coincident locations. Unfortunately, the results from a large number of studies such as those discussed in the aforementioned indicated that the XBT error varied from probe to probe and from year to year. Within a coordinated and international effort, Hanawa et al. (1995) collected hundreds of XBT–CTD pairs with the aim to propose a global estimate of error, which was later confirmed to be insufficient because the errors are time varying (Gouretski and Koltermann 2007).

Recently, Cowley et al. (2013) collected more than 5000 XBT–CTD pairs (side-by-side XBT–CTD comparison dataset) since 1970s to reexamine and recalculate the XBT errors. Based on the results, they proposed a correction scheme (CW13) for the global XBT dataset (containing ~2.2 million XBT profiles from 1966 to 2010). The CW13 correction scheme considers the most popular probe types and is variable with calendar years.

It is still debatable whether the results from thousands of pairs from high-quality XBT–CTD comparisons could be representative of the errors of the millions of global XBT data. Cheng et al. (2014) provided some clues for this question by comparing the errors obtained by both a side-by-side XBT–CTD dataset and a global-scale XBT–CTD comparison dataset, which contains XBT and CTD pairs within 1° spatial distance and a 1-month period. They concluded that the two datasets share similar depth errors but different pure temperature errors prior to 1985. This finding suggests that side-by-side datasets are still insufficient to represent the errors in the historical global XBT dataset, given the fact that there are only ~4000 pairs in the side-by-side dataset. But, the XBT bias is complicated and variable with probe types, temperature, and year.

Research to improve corrections of the bias in the historical global XBT dataset includes various statistical methods, such as comparing XBT profiles with collocated reference profiles, that is, CTD data (Wijffels et al. 2008; Ishii and Kimoto 2009; Levitus et al. 2009; Gouretski and Reseghetti 2010; Hamon et al. 2012; Cheng et al. 2014), and a comparison with General Bathymetric Chart of the Oceans (GEBCO) data (Good 2011; Gouretski 2012). By using these strategies, the number of comparisons between XBTs and reference data is increased to millions (rather than thousands as with the side-by-side dataset), which increases the confidence of the statistics of the size of XBT bias. However, mesoscale signals within a 1° grid and a 1-month period introduce additional uncertainty to the comparison between XBT and CTD. It is assumed that mesoscale noises can be smoothed out by using a large number of data.

Many correction schemes have been proposed from different international groups since 2008 (e.g., Wijffels et al. 2008; Ishii and Kimoto 2009; Levitus et al. 2009; Gouretski and Reseghetti 2010; Hamon et al. 2012; Cheng et al. 2014). This study compares 10 of the most popular schemes that make different assumptions about the sources of bias, use different subsets of global XBT datasets, and different methodologies to detect the bias. Therefore, the existing correction schemes have substantial differences in how they correct the individual XBT profiles. Until now, there are no clear guidelines articulating the respective advantages of one or more of these schemes; thus, there is no general agreement on the best correction scheme.

This study attempts to examine the performance of the 10 existing XBT correction schemes by correcting the XBT data in two major ocean subsurface available databases: the World Ocean Database 2013 (WOD13; Boyer et al. 2013) and the EN4 dataset (Good et al. 2013). Two databases are used because they differ in population size and quality control (QC), which may lead to differences in the XBT correction results. These 10 schemes were selected because they could be readily applied to the global XBT dataset. There is one other global scheme not examined in this study: the altimetry method in Wijffels et al. (2008, their Table 2). A brief summary of these 10 correction schemes is reviewed in Cheng et al. (2016) and a careful intercomparison among those schemes has been recommended in that study.

Two classes of tests were constructed in our analyses. The first one was named the “overall correction” test, where all of the XBT data from both the WOD13 and EN4 dataset were corrected by using the 10 correction schemes separately and then the corrected XBT data were compared with collocated reference data, including CTD, profiling floats (PFL) and Ocean Station Data (OSD, bottle casts). Different metrics were proposed to define the “goodness” of the correction, which helps to evaluate the correction schemes. From the 2.2 million XBTs, only 414 000 (~19%) and 351 000 (~16%) were evaluated from WOD13 and EN4, respectively. The remaining XBT profiles do not have collocated reference profiles.
The overall-correction experiment helps test how well various schemes could correct a subset of historical XBT data (those with collocated reference data). One problem with correction schemes is that there are no independent data used for evaluation. With the criteria for matching pairs used in the present work, the number of pairs has been greatly expanded (~20% of all XBT drops). Therefore, we designed a second test (named the “training/testing” test) where half of XBT–reference pairs in this study were used to train a correction scheme (training dataset) and the other half were used to test the performance of the obtained scheme (testing dataset). In this way, we could test the robustness of the correction schemes by using a fully independent dataset.

Furthermore, a high-quality XBT–CTD side-by-side comparison dataset provides an independent means for examining the XBT bias (Cowley et al. 2013). So, the XBT data in this dataset are also corrected by using 10 schemes and compared with CTD data in order to test the performance of schemes on correcting XBT data in this high-quality dataset.

The remainder of the manuscript is constructed as follows: Data and methods are introduced in section 2. The results of the overall-correction test and the training/testing test are collected and discussed in sections 3 and 4, respectively. A further check based on the XBT–CTD side-by-side dataset is provided in section 5. The main results of this study are summarized in section 6, where the impact of XBT schemes on ocean heat content (OHC) estimates since 1966 is discussed.

2. Data and methods

a. Data

Two databases are employed in this study. The first is the WOD13 from the National Oceanic and Atmospheric Administration’s National Centers for Environmental Information (NOAA NCEI; Boyer et al. 2013). XBT, CTD, OSD, and PFL data were collected to identify WOD13-XBT–reference pairs as follows: for each XBT profile, we searched for CTD, PFL, or OSD profiles within 1° spatial distance and a 1-month time difference, and then the nearest profile in geographical location is set as the reference. In Cheng et al. (2014), tests were made to compare the two choices of the reference profile (the nearest profile, and the average of all collocated profiles) and they found nearly identical results; therefore, in this study the nearest reference profile was used for simplicity.

The second database is the EN4 dataset (EN.4.0.2) published by the Hadley Centre of the Met Office (Good et al. 2013). EN.4.0.2 is used instead of the latest version, EN.4.1.1, since XBT data have not been adjusted for bias correction in EN.4.0.2. The main data source of EN.4.0.2 is the WOD. The EN4 XBT–reference comparison dataset is prepared in the same way as the WOD13-XBT–reference, and CTD, PFL, OSD data are used as reference (as indicated in the EN4 dataset, according to the “s_29_probe_type” in the WOD05 document).

For both datasets, QC flags provided by data producers were used to remove the spurious profiles and temperature measurements. For WOD13, only temperature and depth measurements with “flag = 0” are used, and for EN4, all the data with “flag = 4” are removed. Additional checks are included: (i) There must be at least one measurement in the upper 100 m. (ii) Profiles with fewer than five temperature measurements are not used. (iii) The difference between XBT and reference measurements should be less than 5°C (an empirical choice).

In total, we found 542,743 WOD13-XBT–reference pairs, ~25% of the whole collection of XBT data (~2.2 million) in the WOD13 from 1966 to 2010. After QC processes, there are 413,956 remaining pairs used for analysis in this study (~19%). There are different probe types in the WOD13-XBT–reference dataset, including T7/Deep Blue (DB; 88,969 pairs), deep unknown [unknown probe types with maximum depth deeper than 550 m (DX); 24,239 pairs], shallow unknown [unknown probe types with maximum depth shallower than 550 m (SX); 190,397 pairs], T10 (10,732 pairs), and T5 (9,247 pairs) from the manufacturer Sippican (Fig. 1). There were TSK-T4/T6 (10,925 pairs), TSK-T5 (7,382 pairs), and TSK-T7 (5,880 pairs) from the manufacturer TSK. For the EN4-XBT–reference dataset, there are ~427,134 pairs (~350,549 after QC). There are 91,725 T7/DB, 17,932 DX, 145,544 SX, 67,413 T4/T6, 10,110 T10, 7160 T5, 5921 TSK-T4/T6, 502 TSK-T5, and 4242 TSK-T7 profiles in EN4-XBT–reference dataset (Fig. 1).
Sippican stated originally that the physical characteristics of T4/T6/T7/DB are the same and that the fall rate equation has the same coefficients for these four XBT types, but this does not exactly hold, based on experimental results and independent studies (i.e., Green 1984; Hallock and Teague 1992; Hanawa et al. 1995; Reseghetti et al. 2007; Cowley et al. 2013). To group the profiles with unknown probe types, it is better to consider the maximum (max) depth each probe type can reach. Based on the tests mainly in Mediterranean Sea, the maximum depth for each probe type is as follows: T10 (max = 250 m), T4/T6 (max = 550 m), T7/DB (max = 900 m), Fast Deep (FD; max = 1200 m), and T5–T5/20 (max = 2200 m). Occasionally, T4/T6 probes can reach up to 570–580-m depth and T7/DB to 920–930-m depth [when the Hanawa et al. (1995; H95) fall rate coefficients are used], while T10 probes reach 270-m depth. So, the unknown probes in the range up to 930 m are most likely to be T7/DB, so they are grouped in that category. Unknown-type profiles deeper than 930 m are grouped with T5 probes, although it is possible that some unknown profiles recorded after 1990 could also be FD.

All of the obtained profiles are linearly interpolated to standard depths (1- and 5-m intervals from 5 to 2000 m) during the analyses. Many XBT profiles do not contain information about the applied fall rate equation, so we follow the decision made in the WOD13 dataset assuming that the manufacturer FRE has been used [when “depth.fix” is equal to 1, we correct the original fall rate equation (FRE) to the H95 equation]. EN4 has already been applied this correction according to their data description, so we did not do any corrections for EN4 data. As the GR10 correction scheme requires that the XBT profiles are corrected with the manufacturer FRE, all depth observations (for T4/T6, T7/DB, SX, and DX) are multiplied by a factor of 0.967 prior to applying the GR10 scheme.

b. Method

1) OVERALL-CORRECTION TEST

XBT profiles in the two XBT–reference comparison datasets are corrected by using 10 correction schemes separately listed on the XBT-error website (http://www.nodc.noaa.gov/OC5/XBT_BIAS/xbt_bias.html), including their updated history. The 10 schemes include GR10 (Gouretski and Reseghetti 2010); CH14 (Cheng et al. 2014); Levitus et al. (2009; L09); CW13, CH method in Cowley et al. (2013; CWCH); Hamon et al. (2012; H12); Gouretski (2012; G12); Ishii and Kimoto (2009; IK09); Wijffels et al. (2008; W08); and Good (2011; GD11). The latest available version, CW13, is applied. Among these methods, G11/IK09/GD11 schemes attribute XBT bias to depth error, while CH14/H12/G12/GR10 deal with both temperature and depth error, L09 corrects only the temperatures, and CW13/CWCH correct both errors but based only on side-by-side comparison data. Only CH14 and GR10 take into account the influence of ocean temperature on both depth and temperature error, while H12 takes into account the temperature dependency on temperature error.

In this study we compare the performance of correction schemes at 0–700-m layers, because the amount of data available below 700 m is sharply reduced and some correction schemes do not offer corrections below 700 m. Careful examination of XBT errors below 700 m is required in the future and it is dominated by T5 errors (correction for the T5 probe is available only for some schemes). Moreover, many correction schemes provide only corrections to some most popular probe types, that is, GR10 is only for T4/T6 and T7/DB. No correction to some probe types is the major caveat of many correction schemes as discussed in Cheng et al. (2016). In this study our evaluation makes reasonable assumptions for these correction schemes without a correction for some specific probe types. For example, we always apply T4/T6 corrections to shallow XBTs (with maximum depth less than 550 m) and apply T7/DB corrections to deep XBTs (deeper than 550 m) if there is no correction to a specific probe type: that is, we apply T4/T6 correction of GR10 to T10, SX, TSK-T4; and apply T7/DB correction of GR10 to T5, DX, TSK-T5, and TSK-T7. These practical assumptions make our intercomparison of correction schemes more than “a fair comparison,” because a correction is always better than doing nothing, especially when calculating historical OHC change.

Furthermore, different correction schemes are available for different years: for T4/T6, the start time ranges from 1966 (IK09, L09, CH14) to 1968 (W08, GD11, H12) and the end time ranges from 2005 (W08) to 2010 (IK09, CH14, CW13, CWCH). In this study if there is no correction for a particular year, we decided to use the most recent year with correction available. It is a reasonable assumption because more than 90% of the comparison data are within the common period of 1968–2005, and using only this common period does not change the conclusions of this study.

After correcting XBT data in the WOD13-XBT–reference and EN4-XBT–reference databases by using different correction schemes, it still remains to determine how to define the “goodness” of the corrections. Here, four different metrics are proposed: three metrics consider the average value of the absolute values of the temperature difference between XBT and reference profiles, while the remaining one is based on the standard deviation of the time series of the residuals.
(i) Metric 1: Total XBT bias

We propose that the best method will result in zero residuals (i.e., global mean temperature difference between the XBT profile and the reference profile: \( T_{\text{xbt}} - T_{\text{ref}} \)), which indicates that the total bias in XBT data is fully removed. Therefore, the first metric (metric 1) is defined as the absolute mean of the global-averaged temperature difference profile between XBT and reference data \( \overline{|T_{\text{xbt}}(\text{depth}) - T_{\text{ref}}(\text{depth})|} \). The overbar indicates the area average, which is global average in metric 1. For further details, we first calculate the global mean over all profiles and get the global mean temperature difference profile. Second, we take the absolute mean over depth to get the mean error, as shown in

\[
\text{Metric-1} = \frac{1}{n} \sum_{\text{depth}=1}^{n} |T_{\text{xbt}}(\text{depth}) - T_{\text{ref}}(\text{depth})|, \tag{1}
\]

where there are \( n \) standard levels of depth. The absolute mean over depth is calculated because it possible that some schemes overcorrect at certain depths but undercorrect at some depths; a simple arithmetic average would lead to a compensation over depth. Accordingly, the best scheme should have the smallest value of metric 1.

(ii) Metric 2: Total bias of different probe types

There are various XBT probe types with different designs, so it is still a question of how well various correction schemes can correct the errors of different probe types also due to the experienced different error history. Ideally, the best scheme would remove the total error of each probe type. The second metric is the full-depth average of the absolute values of temperature difference for each XBT probe type as seen in

\[
\text{Metric-2(probe)} = \frac{1}{n} \sum_{\text{depth}=1}^{n} |T_{\text{xbt}}(\text{depth,probe}) - T_{\text{ref}}(\text{depth,probe})|, \tag{2}
\]

It is seen that metric 2 is similar to metric 1 but that it is calculated separately for each probe type. It is possible that a correction scheme works in a different way on different probe types.

(iii) Metric 3: Time variation of XBT bias

It has been well documented that the biases in XBT datasets are variable with time (i.e., calendar year), which could induce spurious long-term trends and interdecadal variability in ocean heat content estimation (Gouretski and Koltermann 2007; Domingues et al. 2008). Therefore, the third metric examines how well different correction schemes remove the time variation of the XBT bias. The best method should result in the smallest time variance. So, the third metric is defined as the standard deviation of the time series of the residuals, which are the annual mean values. For each specific year, the annual mean (arithmetic average) value is calculated by compositing the data within the previous and following two years. A 5-yr window is used here to increase the robustness of the statistics for the yearly XBT bias as examined by a bootstrap test in Cheng et al. (2014),

\[
\text{Metric-3} = \sigma[(T_{\text{xbt}} - T_{\text{ref}})_{\text{year}}]. \tag{3}
\]

If there are no time variations in the XBT biases, then the standard deviation of the temperature difference between the XBT and reference profiles is zero.

(iv) Metric 4: Geographical variation of XBT bias

Another typical feature of XBT biases are that they vary with geography; that is, bias is larger in the low latitudes but smaller in the high latitudes (Gouretski and Reseghetti 2010). We propose that the best scheme should remove the zonal variations of the biases. So, the fourth metric is defined as the mean of the absolute value of temperature differences between the XBT and reference at each 1° latitude (70°S–70°N) and depth (5–700 m) as follows:

\[
\text{Metric-4} = \frac{1}{n} \sum_{\text{depth}=1}^{n} \sum_{\text{latitude}=-70}^{70} |T_{\text{xbt}}(\text{depth,latitude}) - T_{\text{ref}}(\text{depth,latitude})|/140n, \tag{4}
\]

where \( n \) is the number of vertical levels. Here the overbar indicates area average at each latitude.

2) Training/testing test

To investigate the robustness of the correction schemes, 50% of the WOD13-XBT–reference data (training dataset) are selected to train the scheme and the other 50% of the data (testing dataset) are used to test the performance of the obtained scheme. In this way the testing dataset is an independent dataset used only for evaluation, since it is not used to derive the correction scheme. Only two schemes—L09 and CH14—are examined in this section for the following reasons: (i) The two methods show outstanding performance according to our results of the overall-correction test (see section 3). (ii) CH14 consider all influencing factors of XBT errors (see Cheng et al. 2016), while L09 is a simple and widely used method, directly comparing temperatures between XBT and CTD in a global dataset.

The two databases were split by working geographically from the south to north and eastward from the date line, placing the first match in the training dataset, the second match in the testing dataset, etc. The outcome of this method is geographically coherent coverage: the
pairs in the two datasets have a nearly consistent geographical distribution.

3) “SIDE BY SIDE DATASET” TEST

It remains to be seen how well the scheme removes the XBT errors in the side-by-side XBT–CTD comparison datasets, such as presented in Cowley et al. (2013). These side-by-side comparisons are high-quality data, and they avoid the influences of the ocean mesoscale signals by comparing XBT profiles with collocated and nearly simultaneous CTD profiles. In this study we also examine the performance of different schemes by correcting the XBT data in a side-by-side dataset, where 2720 Sippican T4, T6, DB, and T7 pairs are available. Both high-resolution and low-resolution pairs are used in this study.

3. Results of the overall correction test

Using the methods described in the previous section, we correct the XBT data and present the results in this section for four metrics separately (sections 3a–d) and a summary is given in section 3e. An additional section (section 3f) will further discuss why there is a difference in XBT errors between the WOD13- and EN4-based datasets.

a. Metric 1: Total XBT bias

Figure 2a shows the overall difference between uncorrected XBT and reference temperatures as a function of depth based on both the WOD13 and EN4 datasets (in black). The term uncorrected is used here to refer to the standard H95 FRE, since the H95 FRE became a standard in 1995 and most of the reported XBT data have used H95 FRE. Generally, XBT data have a warm bias on a global average of 0.131 ± 0.024°C for WOD13 and 0.124 ± 0.025°C for EN4, with one standard deviation error bar. WOD13 data show a slightly larger error than EN4 over 0–700-m layers (with a 0.01°C difference). This difference is examined in section 3f. Besides, there is a small change in bias within 400–500 m as a result of the difference in the T4/T6 and T7/DB profiles (the maximum depth of T4/T6 is ~550 m and lots of data end at 400–500 m in the early era), where the results are dominated by T4/T6 at the upper 400 m and T7–DB below 500 m.

After correcting XBT profiles using 10 different schemes, the total XBT biases are all significantly reduced at the 95% confidence level (according to a Student’s t test). All schemes reduce the total XBT bias to less than 0.05°C below 300 m. But some methods (IK09, W08, GD11) apparently underestimate the biases in the upper 300 m, leading to large positive residuals after corrections (>0.1°C in the region 0–100 m).

The best scheme should be able to fully remove the XBT biases at each depth, that is, metric 1 is zero, as shown in Fig. 2b and summarized in Table 1. Metric 1 is the absolute average over depth after doing the global average of the residuals, so the error will not compensate over depth. The best method reduces the XBT bias to 0.009 ± 0.014°C (CH14) for WOD13 and 0.009 ± 0.010°C (L09) for EN4 (Table 1). For both datasets, the top two methods (CH14 and L09) are significantly different from the other schemes, ranking from 5 to 10 at the 95% confidence level according to a Student’s t test.

b. Metric 2: Total bias of different probe types

The residuals of temperature differences between XBT and CTD for each probe type are shown in Fig. 3 for WOD13 and Fig. 4 for EN4, and the values of metric 2 are summarized in Table 2. In the table the metric 2 performance is shown for each uncorrected (uncorr) probe. Next, each of the 10 schemes is applied to each probe dataset and metric 2 is calculated. In the next columns, a ranking system is employed that lists the first-, second-, third-, and fourth-best-performing scheme. For instance, for the T7/DB probe type, the top performer using the WOD13 database is CH14, while the top-performing scheme using the EN4 database is L09. The top two schemes are all significantly different from the other schemes with ranks from 5 to 10 (according to a Student’s t test). The second-best performers for this probe type are GR10 and CH14 for the WOD13 and EN4 databases, respectively. Before corrections (black), a systematic warm error is evident for each probe type and again EN4 always shows a slightly smaller error than WOD. However, the mean error for TSK-T5 is always positive in WOD but negative below 200 m in EN; the large difference for TSK-T5 is probably due to the insufficient amount of pairs (<1000 pairs, 738 pairs in WOD and 502 pairs in the EN4 comparison dataset).

After corrections by each scheme, the total biases of different probe types are mostly reduced, but we observe different performances for different probe types, suggesting a lack of a general consensus on the best scheme for all probe types. One possible reason is that these schemes (except CH14) do not provide corrections to all of these probe types, and this study has made some practical assumptions to correct the biases of these probes, which cannot guarantee the performance. However, CH14 is the best method for seven probe types in WOD and three probes in EN (and CH14 is the rank-2 for four groups in EN), suggesting a superior performance than the other methods when the probe type is considered.

c. Metric 3: Time variation of XBT bias

Figure 5 depicts the annual mean of 0–700-m averaged temperature difference for each year between the XBT and reference profiles from 1966 to 2010. Before
correction, the XBT errors show a significant time variation (in black), with a larger positive error (0.15–0.25°C for WOD13) before 1985 and a smaller positive error (0.05–0.15°C for WOD13) since 1990. The leading cause of the bias change in the 1980s is that the recording system changed from a strip chart to a digital system (Cowley et al. 2013). After correction, all of the schemes reduce the time variation of the errors. CH14 is the best method for both WOD13 (0.014°C) and EN4 (0.015°C).

Figures 6, 7 show the residuals of temperature differences at different depths and calendar years for the WOD13 and EN4 databases, respectively. In Fig. 6 it can be seen that the uncorrected situation gives a large prevalent warm error that changes in time and depth over
the past decades. The various corrections methods generally reduce this error and bring its value closer to zero; in some instances, however, errors persist, particularly in shallow layers or very recent years. CH14 and L09 are the best ones for removing the errors, resulting in near-zero (less than 0.05°C) residuals from 0 to 700 m and from 1966 to 2010 (L09 slightly underestimates the errors during 1990s). GR10 and G12 result in a positive residual in the

Table 1. Summary of the values of the three metrics for both the WOD13 and EN4 comparison datasets. One standard deviation error bar is attached.

<table>
<thead>
<tr>
<th>Scheme (°C)</th>
<th>Metric 1</th>
<th>Metric 3</th>
<th>Metric 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WOD13</td>
<td>EN4</td>
<td>WOD13</td>
</tr>
<tr>
<td>Uncor</td>
<td>0.131 ± 0.024</td>
<td>0.124 ± 0.025</td>
<td>0.048</td>
</tr>
<tr>
<td>CH14</td>
<td>0.009 ± 0.014</td>
<td>0.012 ± 0.010</td>
<td>0.014</td>
</tr>
<tr>
<td>W08</td>
<td>0.045 ± 0.050</td>
<td>0.045 ± 0.045</td>
<td>0.042</td>
</tr>
<tr>
<td>IK09</td>
<td>0.067 ± 0.046</td>
<td>0.057 ± 0.048</td>
<td>0.029</td>
</tr>
<tr>
<td>GD11</td>
<td>0.044 ± 0.043</td>
<td>0.047 ± 0.040</td>
<td>0.034</td>
</tr>
<tr>
<td>H12</td>
<td>0.060 ± 0.021</td>
<td>0.048 ± 0.022</td>
<td>0.030</td>
</tr>
<tr>
<td>CWCH</td>
<td>0.043 ± 0.034</td>
<td>0.038 ± 0.032</td>
<td>0.043</td>
</tr>
<tr>
<td>CW13</td>
<td>0.040 ± 0.030</td>
<td>0.031 ± 0.032</td>
<td>0.040</td>
</tr>
<tr>
<td>GR10</td>
<td>0.020 ± 0.025</td>
<td>0.024 ± 0.021</td>
<td>0.019</td>
</tr>
<tr>
<td>L09</td>
<td>0.018 ± 0.010</td>
<td>0.009 ± 0.010</td>
<td>0.022</td>
</tr>
<tr>
<td>G12</td>
<td>0.024 ± 0.025</td>
<td>0.029 ± 0.020</td>
<td>0.030</td>
</tr>
</tbody>
</table>

Fig. 3. Global averaged temperature difference between XBT and CTD profiles from 0 to 700 m for each probe type (curves in the first and third columns). And metric 2: 0–700-m average of the absolute temperature difference for each probe type (dots in the second and fourth columns). The WOD13-XBT-reference dataset is used here.
upper 200 m (up to 0.1°C) from 1980 to 2000 and a negative residual below 200 m especially after 1990. W08/IK09/H12/CWCH/CW13 show consistent warm residuals before 1990, suggesting an underestimation of XBT errors.

d. Metric 4: Geographical variation of XBT bias

The geographical variation of the XBT bias is evident when one looks at the zonal averages of XBT errors in Fig. 8 for WOD13 and Fig. 9 for EN4. In the figures, pairs within each 1° zonal band are averaged together. A larger warm error is evident in low latitudes within 30°S–30°N (0.2°–0.3°C) in the 100–500-m depth interval and a smaller error at high latitudes in the Southern Hemisphere (70°S–40°S). As defined in the previous section, metric 4 reveals the strength of the zonal variation of XBT errors also indicating whether a correlation with the temperature of the water column exists (e.g., Cowley et al. 2013; Cheng et al. 2014; Abraham et al. 2016) or whether the variations of XBT errors are of the same kind of variation of water characteristics with the latitude.

All of the schemes show reduced values of metric 4 compared with the uncorrected data. Before corrections, the zonal absolute mean error (metric 4) is 0.120 ± 0.081°C for WOD13 and 0.116 ± 0.082°C for EN4, while it can best be reduced by CH14 to 0.039 ± 0.056°C and 0.042 ± 0.060°C, respectively. CH14 results are significantly different from the results without correction (Student’s t test) but not significantly different from the other schemes, mainly because the zonal mean temperature differences shown in Figs. 8, 9 are not white noises apparently, so the assumption of a norm distribution in a Student’s t test does not valid. Nevertheless, CH14 is the best one for removing the zonal mean (metric 4) and the variation (error bar of metric 4) of XBT errors, indicating the importance of including a temperature-dependent term in the correction. CH14 slightly overestimates the errors in the Southern Ocean (70°S–40°S, residuals less than −0.05°C). L09 greatly overestimates the errors in the Southern Ocean (residuals range from −0.2° to −0.1°C) but underestimates the errors in low latitudes in the Southern Hemisphere.

FIG. 4. As in Fig. 3, but the EN4-XBT–reference dataset is used.
<table>
<thead>
<tr>
<th></th>
<th>Uncorr (°C)</th>
<th>Rank 1 (scheme, °C)</th>
<th>Rank 2 (scheme, °C)</th>
<th>Rank 3 (scheme, °C)</th>
<th>Rank 4 (scheme, °C)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WOD EN4</td>
<td>WOD EN4</td>
<td>WOD EN4</td>
<td>WOD EN4</td>
<td>WOD EN4</td>
</tr>
<tr>
<td>T7/DB</td>
<td>0.127 ± 0.010 0.117 ± 0.011</td>
<td>CH14 L09</td>
<td>0.004 ± 0.006 0.010 ± 0.008</td>
<td>0.014 ± 0.010 0.011 ± 0.004</td>
<td>0.020 ± 0.006 0.012 ± 0.14</td>
</tr>
<tr>
<td>DX</td>
<td>0.089 ± 0.024 0.078 ± 0.024</td>
<td>CH14 CWCH</td>
<td>0.010 ± 0.015 0.010 ± 0.016</td>
<td>0.011 ± 0.015 0.011 ± 0.013</td>
<td>0.012 ± 0.008 0.012 ± 0.012</td>
</tr>
<tr>
<td>T4/T6</td>
<td>0.177 ± 0.029 0.179 ± 0.029</td>
<td>CH14 L09</td>
<td>0.016 ± 0.017 0.016 ± 0.016</td>
<td>0.026 ± 0.028 0.026 ± 0.028</td>
<td>0.029 ± 0.029 0.030 ± 0.029</td>
</tr>
<tr>
<td>SX</td>
<td>0.145 ± 0.023 0.141 ± 0.022</td>
<td>CH14 L09</td>
<td>0.014 ± 0.008 0.010 ± 0.011</td>
<td>0.025 ± 0.014 0.024 ± 0.015</td>
<td>0.029 ± 0.013 0.024 ± 0.013</td>
</tr>
<tr>
<td>T10</td>
<td>0.070 ± 0.024 0.076 ± 0.024</td>
<td>CWCH CWCH</td>
<td>0.019 ± 0.017 0.019 ± 0.018</td>
<td>0.021 ± 0.017 0.018 ± 0.018</td>
<td>0.026 ± 0.023 0.027 ± 0.027</td>
</tr>
<tr>
<td>T5</td>
<td>0.128 ± 0.010 0.120 ± 0.015</td>
<td>CH14 CH14</td>
<td>0.016 ± 0.007 0.015 ± 0.011</td>
<td>0.017 ± 0.016 0.018 ± 0.018</td>
<td>0.064 ± 0.064 0.053 ± 0.038</td>
</tr>
<tr>
<td>TSK-T4/T6</td>
<td>0.156 ± 0.082 0.125 ± 0.073</td>
<td>CH14 GD11</td>
<td>0.058 ± 0.037 0.046 ± 0.042</td>
<td>0.064 ± 0.064 0.053 ± 0.038</td>
<td>0.071 ± 0.075 0.054 ± 0.043</td>
</tr>
<tr>
<td>TSK-T5</td>
<td>0.124 ± 0.048 0.124 ± 0.091</td>
<td>L09 G12</td>
<td>0.043 ± 0.034 0.090 ± 0.061</td>
<td>0.053 ± 0.031 0.156 ± 0.116</td>
<td>0.061 ± 0.039 0.189 ± 0.101</td>
</tr>
<tr>
<td>TSK-T7</td>
<td>0.050 ± 0.030 0.060 ± 0.032</td>
<td>CH14 CH14</td>
<td>0.014 ± 0.010 0.018 ± 0.015</td>
<td>0.020 ± 0.016 0.021 ± 0.020</td>
<td>0.034 ± 0.028 0.035 ± 0.012</td>
</tr>
</tbody>
</table>

These results indicate that introducing correction schemes (CH14, 2016, i.e., water temperature, probe types, time variations) helps remove errors. CH14 seems to slightly overestimate the errors in the Southern Hemisphere and to slightly underestimate the errors in the Northern Hemisphere (L09). CH14 is the best scheme for almost all of the metrics. The scheme was found to be effective for correcting for the errors in the Southern Hemisphere. CH14 seems to be the best scheme for almost all of the metrics. The scheme was found to be effective for correcting for the errors in the Southern Hemisphere.
as shown in Figs. 8, 9, leaving a large cold error in the southern oceans and a warm error in low latitudes.

- GR10 and G12 are the third- and fourth-best schemes, respectively. GR10 strongly overestimates the errors in low latitudes (30°S–30°N) at the sub-surface (200–700 m; Figs. 8, 9). This is likely the reason why the GR10-based global OHC time series is always cooler in the 1990s than the others using CH14 and L09 (Boyer et al. 2016). The GR10 (and also G12) approach did a fairly good job in removing the time variation of the errors, except for the insufficient correction during the 1980s (mainly in the upper 200 m) and a slight overcorrection since 2000 (below 200 m). Furthermore, we note that G12 (also GD11) is independent of all others, because they rely upon the completely independent information about the bottom depth from the digital bathymetry. This provides independent information on the fall rate error.

- Based on the evaluations shown in sections 3a–e, the recommendation made in Cheng et al. (2016), and the scores in Table 3, we recommend CH14 as the best correction scheme, which is both the best one that takes into account all of the influencing factors of the XBT bias (Cheng et al. 2016) and shows the best performance. We also support the future application of L09/GR10/G12 in scientific studies, with the caveats about geographic problems (for L09 and GR10) and temporal problems (for GR10 and G12). Despite their relatively poor performance (compared to other schemes), we recommend that both CW13 and CWCH be maintained.

![Figure 5](image-url)
in the future. They are unique in that they contained high-quality side-by-side measurement information and, in principle, such information can provide accurate assessment of XBT errors despite a small volume of data. Also, these two schemes perform well after 2000, and whether this behavior continues is of interest to the community. The schemes of H12/W08/IK09/GD11 are not recommended because of both their insufficiency in taking into account the influencing factors of XBT errors and their fair performances in correcting global XBT errors.

f. Differences between WOD- and EN4-based pairs dataset

The results in the previous sections indicate a difference in XBT errors between the WOD- and EN4-based comparison datasets (i.e., Fig. 2; Tables 1–3), with a global and 0–700-m mean difference of \(-0.01^\circ C\). The difference between the two datasets shown in the previous sections could be caused by three reasons: 1) Different populations in the two datasets, that is, the WOD13-pairs dataset contains more pairs than the EN4-pairs dataset (413,956 vs 350,549); 2) different QC processes, resulting in different definitions of “bad” profiles and bad measurements. These profiles and measurements are removed in our analysis. But quantification of the impact of removing the bad profiles of XBT errors is trivial, since it is related to the statistical issue, that is, subsampling from a population will always lead to a difference in the statistical mean, which is not necessarily caused by QC; and 3) difference in the data processing procedure by the two data centers (e.g., thinning of profiles that contain more than 400 points to 1 m apart above 100 m and 10 m apart to 1500 m). To attempt to identify the reasons for the difference, two tests are now described.

1) To assess the effects of data processing and QC flags (reasons 1 and 3 above), we use the same XBT–reference pairs in the two datasets (identified by the XBT and reference-profile cast numbers), so two subsets are formed: WOD13 pairs with WOD-QC flags (WOD\text{qc}) and EN4 pairs with EN-QC flags (EN\text{ENqc}). The EN4 XBT temperature bias is then subtracted from the WOD XBT temperature bias and the results are shown in Fig. 10.

Figure 10a indicates a median difference of XBT errors of \(-0.005 \pm 0.004^\circ C\) between WOD\text{qc} and EN\text{ENqc} on 0–700-m average (the mean is \(-0.009^\circ C\)). This difference is positive in the upper 800 m (an increase from
approximately $-0.002^\circ C$ near the surface to $-0.01^\circ C$ at 700 m (Fig. 10b), which is evident with or without applying the XBT corrections. This partly (at least 70%: $0.005^\circ C$ compared with a $0.007^\circ C$ difference for metric 1 for the two datasets, which is not significant) explains why the results in the previous sections always show smaller XBT errors for the EN4-pairs dataset compared with the WOD13-pairs dataset (Table 1; Fig. 2). This error might be important in detecting the deep ocean change, because the magnitude of temperature change decreases in the deep ocean (the signal-to-noise ratio decreases with depth).

2) The second test isolates the impact of QC flags from each dataset on XBT bias. We use a set of pairs from the WOD-XBT–reference dataset but apply WOD and EN4 QC flags separately, so two subsets are formed: WODqc and WOD with EN-QC flags (denoted as WODENqc). In this way, the differences in the XBT errors in the two subsets can be purely attributed to the difference in QC of removing measurements. Figure 10c shows a very small difference in XBT errors between WODqc and WODENqc on 0–700-m average (with a median of $-0.001^\circ C$). In the upper 50 m, the maximum difference is $-0.02^\circ C$ (median, WODqc $-$ WODENqc) and it increases to $-0.004^\circ C$ at around 80 m (Fig. 10d). It should be noted that these reported uncertainty values are less than the absolute accuracy of the instrument, and while the differences are meaningful, a statistical significance test has not been performed. We can conclude that the application of QC from each dataset contributes a small part to the difference in the XBT bias between WOD and EN4.

The small differences in the datasets as a result of QC can cause differences in the statistics of the XBT bias. For instance, some XBT data contain errors as result of wire leakage, which is always warm. Differences in the QC procedures to remove the warm leakage errors from a population distribution could be responsible for the statistical metrics (mean, median, skewness etc.). A solution to these problems is offered by the ongoing international project the International Quality-Controlled Ocean Database (IQuOD), which provides an opportunity to explore the best practice for QC.

4. Results of training/testing analysis

In the first step of the training analysis, each XBT scheme was trained on the training dataset, and the results were labeled as the trained schemes (i.e., trained CH14, trained L09, etc.). Figure 11 compares the trained and original CH14 for all XBT types and shows the
The trained scheme is generally consistent with the original one, suggesting that even if half of the comparison dataset is used, a robust scheme can be generated. The original and trained CH14 show larger differences for coefficient A in the DX group after 1995 and in the T4/T6 group within 1995–2000, probably because the data amount during these periods is very small for the two groups: T4/T6 probes dominate the XBT dataset prior to 1995 and DX probes are negligible after 1995. For L09, the trained L09 is also very similar to the original one, but it generally shows a larger bias than the original one (Fig. 12). To demonstrate whether there is any sensitivity of the results to data selection within the same dataset, we also trained CH14 and L09 by using the testing dataset (Figs. 11, 12). The CH14 trained by two different datasets show very similar results, suggesting the robustness of the scheme. Some uncertainty occurs when the data amount is small (SX, DX, T10 after 2000; T5 before 1980; TSK-T7 before 2000). Similarly, for L09, nearly identical results can also be found when training this scheme using the testing dataset (Fig. 12). Therefore, switching the testing and training datasets to train CH14 and L09 increases the robustness of the correction schemes.

Are the trained schemes capable of correcting the systematic XBT errors of each probe type? Figures 13a–c compare the temperature differences between XBT and the reference profiles in the training datasets before and after correction using the trained schemes. Before correction, we see warm biases ranging from 0.04°C to 0.35°C for different probe type groups, but after correction with the trained CH14 (Fig. 13b) and the trained L09 (Fig. 13c), the warm biases are largely removed with residuals within ±0.04°C (except TSK-T5, TSK-T4/T6), indicating that the XBT bias is successfully quantified in the training dataset. The total bias in the
training data is reduced to within ±0.02°C. A large fluctuation of bias occurs for TSK-T5 data because there are fewer than 1000 pairs for this type. Biases in the T10 and the TSK-T4/T6 probe types are reduced but not fully eliminated by both trained schemes, since the temperature differences change over depth. This implies that their bias may have not been fully modeled by the existing schemes (further careful examination is needed). The trained-L09 scheme underestimates TSK-T5 by ~0.15°C and TSK-T4/T6 by ~0.10°C because L09 does not take into account the differences among probe types.

Next, we applied the trained schemes to correct the XBT profiles in the testing dataset to check whether the trained schemes are also valid in this independent group. The results are shown in Figs. 13d–f. All of the main probe types (T7/DB, DX, T4/T6, SX, T5, TSK-T7)

Table 3. Scores of different correction schemes according to four different metrics proposed in this study. The score is defined as the rank of the performance of the schemes for each metric. The rows in bold indicate the four best schemes and the rows in italics denote the schemes based on side-by-side data.

<table>
<thead>
<tr>
<th>Metric 1 total bias</th>
<th>Metric 2 type</th>
<th>Metric 3 time variation</th>
<th>Metric 4 latitude</th>
<th>Total score</th>
<th>Total score</th>
<th>Total score</th>
</tr>
</thead>
<tbody>
<tr>
<td>WOD</td>
<td>EN4</td>
<td>WOD</td>
<td>EN4</td>
<td>WOD</td>
<td>EN4</td>
<td>WOD</td>
</tr>
<tr>
<td>CH14</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>L09</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>GR10</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>G12</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>CW13</td>
<td>5</td>
<td>5</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>H12</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>CWCH</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>W08</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>GD11</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>IK09</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Fig. 9. As in Fig. 8, but EN4 data are used (°C).
show significantly reduced errors after correction compared with the reference profiles (residuals within \( \pm 0.04^\circ \text{C} \)). This independent test reveals that the trained scheme is effective in removing the XBT bias. The T10 and TSK-T4/T6 probes show similar depth variation as in the training dataset. TSK-T5 is the only exception, although its systematic positive error is reduced.

To summarize the observations, the training and testing experiments indicate that the two schemes—L09 and CH14—can still effectively detect and reduce the majority of XBT errors even when there are only half the data. This suggests that the schemes are robust in detecting XBT errors.

5. Results of side-by-side XBT–CTD dataset

In this section the XBT correction schemes are applied to the side-by-side XBT–CTD comparison dataset. Global results can be visualized in Figs. 14, 15. It is worthy to note at the beginning that we are not using the side-by-side dataset to rank and evaluate the correction schemes. There are many caveats in side-by-side data: 1) There are only 4000 pairs, which is an insufficient amount for representing XBT bias that is impacted by many influencing factors (probe types, temperature, time, geography). 2) In the side-by-side data, only T4/T6, T7/DB, and a few TSK (<200) pairs are available. 3) There are no unknown types in side-by-side datasets, but nearly half of the global-scale data are composed of an unknown probe type. Because of these caveats, we are not to use side-by-side data to evaluate the schemes. However, a side-by-side dataset is a scientifically quality-controlled dataset (i.e., all bad data are removed), and the comparisons are much less impacted by the ocean variability between XBT and collocated CTD profiles compared with the global-scale dataset. So, the community aims to highlight and maintain the side-by-side dataset because it is a high-quality comparison, and it is a valuable tool to understand the influencing factors of XBT bias. For example, CH14 uses side-by-side data to get the correlation among three fall rate coefficients, and obtains the correlation between the fall rate (pure temperature error) and ocean temperature. And then CH14 uses global-scale data to get the time variation of both the fall rate and pure temperature error.
error. Therefore, a mixed XBT scheme (i.e., CH14) could combine the best of these two datasets to help reduce biases.

For T4/T6 probes, IK09 is the best at removing both the depth error (Fig. 14a) and the total temperature error below 100 m (Fig. 15a). L09 does not correct the depth error, but it is one of the best ones for removing the total temperature error. W08/G12/GD11/GR10 schemes overcorrect the depth error and induce negative temperature errors in side-by-side datasets (residual temperature differences range from $-0.12^\circ$ to $-0.05^\circ$C). But CH14 slightly underestimates the depth error (4 m at 400 m, 1% error) and the residual temperature error is negative ($-0.05^\circ$C; Fig. 15a). This behavior indicates that pure temperature error plays a dominant role in the total error of T4/T6, confirming the difference between side-by-side data and global-scale data in Cheng et al. (2014). Most of the Sippican T4/T6 data were collected.

Fig. 11. (a)–(i) Comparison between trained CH14 using the training dataset (red) and the original CH14 scheme (black) for nine different probe type groups. The switched trained CH14 using the testing dataset is also shown (blue). Both coefficient $A$ and the pure temperature error are shown.
on analog recorders (pre-1990), which are shown to have a significantly higher temperature bias than digital recorders (Cowley et al. 2013). The four best-performing schemes for the side-by-side datasets (the smaller the temperature residual, the better the scheme) are CW13, CWCH, L09, and IK09. However, for global-scale data, the four schemes are CH14, L09, G12, and GR10.

For T7/DB probes, L09 can fully remove the total temperature errors from 200 to 700 m (Fig. 15b), although it does not correct depth error. CH14 is similar to CWCH in that it can reduce the depth error (less than 2 m; Fig. 14b), but the residual temperature differences are reduced to approximately $-0.03^\circ$C on 0–750-m averages (similar to CWCH again). GR10 greatly overestimates the depth error (residual depth error of $-5$ m at 700 m).

![Comparison between (a) original L09 scheme, (b) trained L09, and (c) switched trained L09 using the testing dataset.](image-url)

**FIG. 12.** Comparison between (a) original L09 scheme, (b) trained L09, and (c) switched trained L09 using the testing dataset.

![Temperature difference between the XBT and reference data in the (top) training and (bottom) testing datasets: (a),(d) without any corrections; (b),(e) applied trained-CH14 correction, and (c),(f) applied trained-L09 correction.](image-url)

**FIG. 13.** Temperature difference between the XBT and reference data in the (top) training and (bottom) testing datasets: (a),(d) without any corrections; (b),(e) applied trained-CH14 correction, and (c),(f) applied trained-L09 correction.
FIG. 14. Comparison of mean residual depth errors (XBT minus CTD) in the side-by-side dataset after depth corrections from the CW13, CWCH, CH14, GR10, GD11, W08, G12, IK09, and H12 methods are applied. Shown are results for Sippican (a) T4/T6 and (b) T7/DB probe types. The “raw” label indicates uncorrected data.
Fig. 15. Comparison of mean residual temperature errors (XBT minus CTD) in the side-by-side dataset after depth corrections from the CW13, CWCH, CH14, GR10, GD11, W08, G12, IK09, H12, and L09 methods are applied. Shown are results for Sippican (a) T4/T6 and (b) T7/DB probe types. The “raw” label indicates uncorrected data.
which results in the largest residual temperature differences (approximately 0.08°C). The four best schemes for side-by-side datasets are CW13, L09, CWCH, and H12. However, for global-scale data, they are CH14, L09, GR10, and G12.

The difference in performance of the schemes on the global datasets and the high-quality XBT–CTD pairs datasets can be attributed to how the schemes are derived, the data quality in the datasets, and the difference in data amount. The CW13 and CWCH schemes were derived by this side-by-side dataset, so it is not surprising that the two schemes show better performance than the others. However, these two schemes are based on a side-by-side dataset. The two schemes provide useful and independent information for XBT errors, so we recommend maintaining the two schemes in the future. Reconciliation of the two major datasets is an important follow-up research topic.

Additionally, the CW13 and CWCH schemes do not provide a correction for T10s or T5s, but they provide a very poor correction for TSK probes because of the extremely low number of TSK XBT–CTD pairs in the pairs dataset. In this study CW13 and CWCH Sippican T4/T6 corrections were applied to all T10, T11, and unknown types with a terminal depth < 550 m, whereas Sippican T7/DB corrections were applied to Sippican Fast Deeps and unknown types with terminal depth ≥ 550 m. However, this approach is not adequate for FD and T5, because if a profile has value at a depth deeper than 900–930 m, then it should be an FD or T5 probe (most likely an T5, since there are more T5 used in history). CW13 and CWCH did not provide a correction for T5, so we still apply this strategy in our comparison.

6. Discussion and conclusions

Attempts were made in this study to compare the performance of 10 among all the existing correction schemes for removing systematic XBT errors from historical XBT data. Three separate experiments were constructed: the first one is an “overall correction” test that corrects all of the XBT data in two reference datasets: WOD13- and EN4-XBT–reference; the other one is a “training/testing” test that corrects half of the XBT data in the WOD13-XBT–reference dataset and leaves the remaining 50% of data as an independent dataset for evaluation. Finally, the correction schemes were applied to an independent XBT–CTD dataset.

According to the results, it is found that

(i) All of the correction schemes substantially reduce the errors in the historical XBT datasets. The performance of different correction schemes varies for different aspects (different metrics). Evidence indicates that CH14 (recommended), L09, GR10, and G12 are the best-performing schemes. The schemes of CH14 and L09 are robust according to training/testing analysis.

(ii) Most of the schemes reduce the errors in the side-by-side dataset, but the performance is not consistent with a global-scale dataset, indicating that more work is required to reconcile the two datasets.

It is also important to note that the issue of XBT bias has not been fully solved yet, since

(i) There are still errors in the upper 100 m after corrections; these remaining errors appear to be positive on a global scale (Fig. 2), negative in the Southern Hemisphere, and positive in the Northern Hemisphere (Fig. 6), mostly for T4/T6 and SX (shallow unknown) probes (Fig. 4). Careful examination of the errors in the upper 100 m is required in the future. One potential approach is via numerical simulation of the probe falling (Abraham et al. 2012a,b; Stark et al. 2011) and via more tests using high-speed photography (Shepard et al. 2014; Schwalbach et al. 2014; Bringas and Goni 2015).

(ii) There are large differences in T10 and T5 biases between the training dataset and the testing dataset, and poor performance for these probe types in most correction schemes. It is likely that this is because there are small sample sizes for the two probe types. Therefore, a robust estimate for T10 and T5 bias
requires more data, and it could be enhanced by recovering more historical side-by-side comparisons (iii) QC and data processing could lead to a systematic error in XBT. Further work in understanding the impacts of QC and data processing on deriving XBT biases and other applications (i.e., OHC change) is required. It is also important to provide guidelines for best practices for QC based on community efforts (via the IQuOD project). It is worth noting that different correction schemes use different datasets, for instance, the GR10 collocated dataset is several years older than CH14. Therefore, QC and data processing could be partly responsible for the difference between the correction schemes.

A more comprehensive summary of the existing problems of XBT errors can be found in Cheng et al. (2016). Fully addressing the existing problems will lead to a more preferable correction scheme in the future.

Furthermore, one of the most important applications of XBT data is estimating historical OHC changes (Lyman et al. 2010; Rhein et al. 2013; Boyer et al. 2016; Cheng et al. 2015, 2017; Wang et al. 2017), which is a vital sign of global climate change. It will be interesting to check how large uncertainty will be in global OHC time series if we use the best-performing schemes. Figure 16 shows two global OHC time series in the upper 700 m when using three XBT schemes (CH14 and L09) separately but the same mapping method/quality control/climatology are applied according to Cheng et al. (2017). CH14 and L09 show very similar OHC 0–700-m time series for most years during 1968–2016; the maximum difference is $\pm 2.0 \times 10^2$ J during 1995–2001. The largest error is found during 1990–2001 because of the dominance of XBT data during this period. The similarity of the two top schemes (L09 and CH14), with the mean standard deviation of only $0.31 \times 10^2$ J for 1970–2004, is much smaller than that in Boyer et al. (2016), which is $1.21 \times 10^2$ J. This implies that it is potentially capable of refining the uncertainty in the OHC estimate caused by XBT errors in the future. The similarity is likely to depend on the mapping method (Boyer et al. 2016), so a more comprehensive examination on the impact of XBT bias on the OHC estimate should be combined with the mapping method.

Acknowledgments. This work is supported by the National Key Research and Development Program of China (2017YFA0603202, 2016YFC1401705), the Chinese government (315030401), and the National Natural Science Foundation of China (Grant 41476016). We thank Catia Domingues for her valuable comments on this study. We acknowledge NOAA NCEI for providing various XBT corrections, which could be easily downloaded from the NCEI website (http://www.nodc.noaa.gov/OC5/SELECT/dbsearch/dbsearch.html). The Met Office in the United Kingdom is also acknowledged for publishing its QC temperature data. And we acknowledge Simon Good for his help on the EN4 data. The side-by-side dataset can be downloaded in CSIRO data protocol (http://doi.org/10.4225/08/52AE99A4663B1). All of the data (WOD13-XBT-reference dataset and EN4-XBT-reference dataset) used in this study are freely available (http://159.226.119.60/cheng/).

REFERENCES


